Performance of Machine Learning and Big Data Analytics Paradigms in Cyber-security and Cloud Computing Platforms

Gabriel Kabanda
Zimbabwe Academy of Sciences, TREP Building, University of Zimbabwe, Harare, Zimbabwe

Keywords: Cybersecurity, Artificial Intelligence, Machine Learning, Deep Learning, Big Data Analytics, Cloud Computing.

Abstract: The purpose of the research is to evaluate Machine Learning and Big Data Analytics paradigms for use in Cybersecurity. Cybersecurity refers to a combination of technologies, processes and operations that are framed to protect information systems, computers, devices, programs, data and networks from internal or external threats, harm, damage, attacks or unauthorized access. The main characteristic of Machine Learning (ML) is the automatic data analysis of large data sets and production of models for the general relationships found among data. ML algorithms, as part of Artificial Intelligence, can be clustered into supervised, unsupervised, semi-supervised, and reinforcement learning algorithms. The Pragmatism paradigm, which is in congruence with the Mixed Method Research (MMR), was used as the research philosophy in this research as it epitomizes the congruity between knowledge and action. The researcher analysed the ideal data analytics model for cybersecurity which consists of three major components which are Big Data, analytics, and insights. The information that was evaluated in Big Data Analytics includes a mixer of unstructured and semi-structured data including social media content, mobile phone records, web server logs, and internet click stream data. Performance of Support Vector Machines, Artificial Neural Network, K-Nearest Neighbour, Naive-Bayes and Decision Tree Algorithms was discussed. To avoid denial of service attacks, an intrusion detection system (IDS) determined if an intrusion has occurred, and so monitored computer systems and networks, and then raised an alert when necessary. A Cloud computing setting was added which has advanced big data analytics models and advanced detection and prediction algorithms to strengthen the cybersecurity system. The research results presented two models for adopting data analytics models to cybersecurity. The first experimental or prototype model involved the design, and implementation of a prototype by an institution and the second model involved the use service provided by cloud computing companies. The researcher also demonstrated how this study addressed the performance issues for Big Data Analytics and ML, and its impact on cloud computing platforms.

1 INTRODUCTION

1.1 Background

The era of the Internet of Things (IoT) generates huge volumes of data collected from various heteregenous sources which may include mobile devices, sensors and social media. This Big Data presents tremendous challenges on the storage, processing and analytical capabilities. Cloud Computing provides a cost-effective and valid solution in support of Big Data storage and execution of data analytic applications. IoT requires both cloud computing environment to handle its data exchange and processing; and the use of artificial intelligence (AI) for data mining and data analytics. A hybrid cybsecurity model which uses AI and Machine Learning (ML) techniques may mitigate against IoT cyber threats on cloud computing environments. Security issues related to virtualisation, containerization, network monitoring, data protection and attack detection are interrogated whilst strengthening AI/ML security solutions that involve encryption, access control, firewall, authentication and intrusion detection and prevention systems at the appropriate Fog/Cloud level.

Cybersecurity consolidates the confidentiality, integrity, and availability of computing resources, networks, software programs, and data into a

Kabanda, G.
Performance of Machine Learning and Big Data Analytics Paradigms in Cyber-security and Cloud Computing Platforms.
DOI: 10.5220/0010789900003167
In *Proceedings of the 1st International Conference on Innovation in Computer and Information Science (ICICIS 2021)*, pages 33-50
ISBN: 978-989-758-577-7

coherent collection of policies, technologies, processes, and techniques to prevent the occurrence of an attack (Berman et al., 2019). Cybersecurity refers to a combination of technologies, processes and operations that are framed to protect information systems, computers, devices, programs, data and networks from internal or external threats, harm, damage, attacks or unauthorized access (Sarker et al., 2020).

The Network Intrusion Detection Systems (NIDS) is a category of computer software that monitors system behaviour with a view to ascertain anomalous violation of security policies and distinguishes between malicious users and the legitimate network users (Bringas and Santos, 2010). According to (Truong et al., 2020), the components in Intrusion Detection and Prevention Systems (IDPSs) can be sensors or agents, servers, and consoles for network management. An intrusion detection and prevention system (IDPS), shown on Figure 1 below, is placed inside the network to detect possible network intrusions and, where possible, prevent the cyber attacks. The key functions of the IDPSs are to monitor, detect, analyze, and respond to cyber threats.

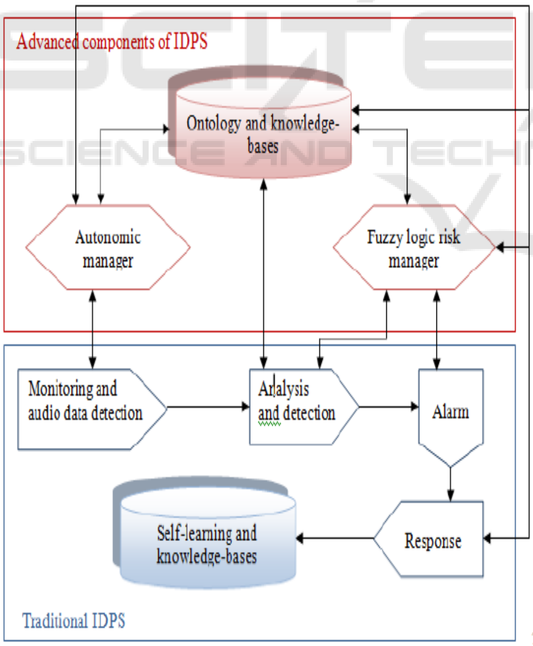

Advanced components of IDPS
Ontology and knowledge-bases
Autonomic manager
Fuzzy logic risk manager
Monitoring and audio data detection
Analysis and detection
Alarm
Self-learning and knowledge-bases
Response
Traditional IDPS

Figure 1: Typical Intrusion detection system.

Computers are instructed to learn through the process called Machine Learning (ML), a field within artificial intelligence (AI). The main characteristic of ML is the automatic data analysis of large data sets and production of models for the general relationships found among data. ML algorithms require empirical data as input and then learn from this input. Deep Learning (DL), as a special category of ML, brings us closer to AI. The three classes of ML are as illustrated on Figure 2 below [5], and these are:

a) ***Supervised Learning:*** where the methods are given inputs labeled with corresponding outputs as training examples;

b) ***Unsupervised Learning***: where the methods are given unlabeled inputs;

c) ***Reinforcement Learning:*** where data is in the form of sequences of observations, actions, and rewards.

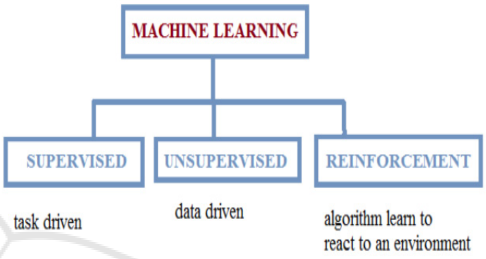


Figure 2: Three levels of Machine Learning (Source: (Proko et al., 2018)).

The transformation and expansion of the cyber space has led to the generation, use, storage and processing of big data, that is, large, diverse, complex, multidimensional and usually multivariate datasets (Mazumdar and Wang, 2018). (Hashem et al., 2015) explained big data as the increase in volume of data that offers difficulty in storage, processing and analysis through the traditional database technologies. Big Data came into existence when the traditional relational database systems were not able to handle the unstructured data generated by organizations, social media, or from any other data generating source (Mahfuzah et al., 2017). Big data analytics makes use of analytic techniques such as data mining, machine learning, artificial learning, statistics, and natural language processing. In an age of transformation and expansion in the Internet of Things (IoT), cloud computing services and big data, cyber-attacks have become enhanced and complicated (Moorthy et al., 2014), and therefore cybersecurity events become difficult or impossible to detect using traditional detection systems (Cox and Wang, 2014), (Hammond, 2015). Big Data has also been defined according to the 5Vs as stipulated by (Yang et al., 2017) where:

❖ Volume refers to the amount of data gathered and processed by the organisation

❖ Velocity referring to the time required to do processing of the data

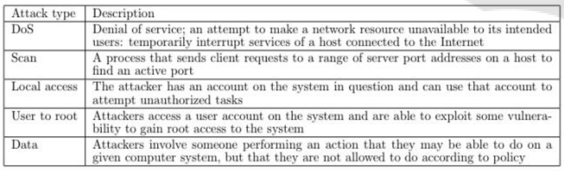
❖ Variety refers to the type of data contained in Big Data

❖ Value referring to the key important features of the data. This is defined by the added-value that the collected data can bring to the intended processes.

❖ Veracity means the degree in which the leaders trust the information to make a decision.

The supervised machine learning algorithm which can be used for both classification or regression challenges is called the Support Vector Machine (SVM). The original training data can be transformed into a higher dimension where it becomes separable by using the SVM algorithm which searches for the optimal linear separating hyperplane. The easiest and simplest supervised machine learning algorithm which can solve both classification and regression problems is the k-nearest neighbors (KNN) algorithm. Both the KNN and SVM can be applied to finding the optimal handover solutions in heterogeneous networks constituted by diverse cells. The Hidden Markov Model (HMM) is a tool designed for representing probability distributions of sequences of observations. The list of supervised learning algorithms includes Regression models, K-nearest neighbors, Support Vector Machines, and Bayesian Learning (Jiang et al., 2016). In Table 1, we summarize the basic characteristics and applications of supervised machine learning algorithms.

Table 1: Various attack descriptions (Source: (Mazumdar and Wang, 2018)).



Attack type	Description
DoS	Denial of service; an attempt to make a network resource unavailable to its intended users: temporarily interrupt services of a host connected to the Internet
Scan	A process that sends client requests to a range of server port addresses on a host to find an active port
Local access	The attacker has an account on the system in question and can use that account to attempt unauthorized tasks
User to root	Attackers access a user account on the system and are able to exploit some vulnerability to gain root access to the system
Data	Attackers involve someone performing an action that they may be able to do on a given computer system, but that they are not allowed to do according to policy

1.2 Statement of the Problem

Firewall protection has proved to be inadequate because of gross limitations against external threats. The rapid development of computing and digital technologies, the need to revamp cyber defense strategies has become a necessity for most organisations (Proko et al., 2018). As a result, there is an imperative for security network administrators to be more flexible, adaptable, and provide robust cyber defense systems in real-time detection of cyber threats.The key problem is to evaluate Machine Learning (ML) and Big Data Analytics (BDA) paradigms for use in Cybersecurity.

1.3 Purpose of Study

The research is purposed to evaluate Machine Learning and Big Data Analytics paradigms for use in Cybersecurity.

1.4 Research Objectives

The research objectives are to:

1)Evaluate Machine Learning and Big Data Analytics paradigms for use in cybersecurity.

2)Develop a Cybersecurity system that uses Machine Learning and Big Data Analytics paradigms.

1.5 Research Questions

The main research question was:

Which Machine Learning and Big Data Analytics paradigms are most effective in developing a Cybersecurity system?

The sub questions are:

1)How are the Machine Learning and Big Data Analytics paradigms used in Cybersecurity?

2)How is a Cybersecurity system developed that uses Machine Learning and Big Data Analytics paradigms?

2 LITERATURE REVIEW

2.1 Overview

Computers, phones, internet and all other information systems developed for the benefit of humanity are susceptible to criminal activity (Cox and Wang, 2014). Cybercrimes consist of offenses such as computer intrusions, misuse of intellectual property rights, economic espionage, online extortion, international money laundering, non-delivery of goods or services, etc. (Yang et al., 2017). Intrusion detection and prevention systems (IDPS) include all protective actions or identification of possible incidents, and analysing log information of such incidents (Truong et al., 2020). (Yang et al., 2017) recommends the use of various security control measures in an organisation. Various attack descriptions from the outcome of the research by (Mazumdar and Wang, 2018) are shown on Table 1. The monotonic increase in an assortment of cyber threats and malwares amply demonstrates the inadequacy of the current countermeasures to defend computer networks and resources. To alleviate the problems of classical techniques of cyber security, research in artificial intelligence and more specifically machine learning is sought after (Berman et al., 2019), (Sarker et al., 2020). To enhance the malware and cyber-attack detection rate, one can apply deep learning architectures to cyber security.

2.2 Classical Machine Learning (CML)

Machine Learning (ML) is a field in artificial intelligence where computers learn like people. We present and briefly discuss the most commonly used classical machine learning algorithms.

2.2.1 Logistic Regression (LR)

As an idea obtained from statistics and created by (Petrenko and Makovechuk, 2020), logistic regression is like linear regression, yet it averts misclassification that may occur in linear regression. Unlike linear regression, logistic regression results are basically either ‘0’ or ‘1’. The efficacy of logistic regression is mostly dependent on the size of the training data.

2.2.2 Naive Bayes (NB)

Naive Bayes (NB) classifier is premised on the Bayes theorem which assumes independence of features. The independence assumptions in Naive Bayes classifier overcomes the curse of dimensionality.

2.2.3 Decision Tree (DT)

A Decision tree has a structure like flow charts, where the root node is the top node and a feature of the information is denoted by each internal node. The algorithm might be biased and may end up unstable since a little change in the information will change the structure of the tree.

2.2.4 K-Nearest Neighbor (KNN)

K-Nearest Neighbor (KNN) is a non-parametric approach which uses similarity measure in terms of distance function classifiers other than news cases. KNN stores the entire training data, requires larger memory and so is computationally expensive.

2.2.5 Ada Boost (AB)

Ada Boost (AB) learning algorithm is a technique used to boost the performance of simple learning algorithms used for classification. Ada Boost constructs a strong classifier using a combination of several weak classifiers. It is a fast classifier and at the same time can also be used as a feature learner. This may be useful in tasks that use imbalanced data analysis.

2.2.6 Random Forest (RF)

Random forest (RF), as an ensemble tool, is a decision tree derived from a subset of observations and variables. The Random Forest gives better predictions than an individual decision tree. It uses the concept of bagging to create several minimal correlated decision trees.

2.2.7 Support Vector Machine (SVM)

Support Vector Machine (SVM) belongs to the family of supervised machine learning techniques, which can be used to solve classification and regression problems. SVM is a linear classifier and the classifier is a hyper plane. It separates the training set with maximal margin. The points near to the separating hype plane are called support vectors and they determine the position of hyper plane.

2.3 Modern Machine Learning

Deep learning is a modern machine learning which has the capability to take raw inputs and learns the optimal feature representation implicitly. This has performed well in various long standing artificial intelligence tasks (Bringas and Santos, 2010). Most commonly used deep learning architectures are discussed below in detail.

2.3.1 Deep Neural Network (DNN)

An artificial neural network (ANN) is a computational model influenced by the characteristics of biological neural networks. The family of ANN includes the Feed forward neural network (FFN), Convolutional neural network and Recurrent neural network (RNN). FFN forms a directed graph in which a graph is composed of neurons named as mathematical unit. Each neuron in i^{th} layer has connection to all the neurons in $i + 1^{th}$ layer.

Each neuron of the hidden layer denotes a parameter *h* that is computed by

$$h_i(x) = f\,(w_iT\ x + b_i) \tag{1}$$

$$hii: Rdi{-}1 \rightarrow Rdi \tag{2}$$

$$f : R \rightarrow R \tag{3}$$

where $w_i \in Rd \times d\ i-1$, $b_i \in Rd_i$, d_i denotes the size of the input, f is a non-linear activation function, ReLU.

The traditional examples of machine learning algorithms include Linear regression, Logistic regression, Linear discriminant analysis, classification and regression trees, Naïve bayes, K-Nearest Neighbour (K-NN), Kmeans clustering Learning Vector Quantization (LVQ), Support Vector Machines (SVM), Random Forest, Monte Carlo, Neural networks and Q-learning. Take note that:

❖ Supervised Adaptation is carried out in the execution of system at every iteration.

❖ Unsupervised Adaptation follows trial and error method. Based on the obtained fitness value, computational model is generalized to achieve better results in an iterative approach.

2.3.2 The Future of AI in the Fight against Cybercrimes

Experiments showed that NeuroNet is effective against low-rate TCP-targeted distributed DoS attacks. (Fernando and Dawson, 2009) presented the Intrusion Detection System using Neural Network based Modeling (IDS-NNM) which proved to be capable of detecting all intrusion attempts in the network communication without giving any false alerts (Menzes et al., 2016).

The characteristics of NIC algorithms are partitioned into two segments such as swarm intelligence and evolutionary algorithm. The Swarm Intelligence-based Algorithms (SIA) are developed based on the idea of collective behaviours of insects in colonies, e.g. ants, bees, wasps and termites. Intrusion detection and prevention systems (IDPS) include all protective actions or identification of possible incidents and analysing log information of such incidents (Truong et al., 2020).

2.4 Big Data Analytics and Cybersecurity

Big Data Analytics requires new data architectures, analytical methods, and tools. Big data environments ought to be magnetic, which accommodates all heterogeneous sources of data. Instead of using mechanical disk drives, it is possible to store the primary data-base in silicon-based main memory, which improves performance. Forecast analytics attempt to predict cybersecurity events using forecast analytics models and methodologies (Petrenko and Makovechuk, 2020). Threat intelligence helps to gather threats from big data, analyze and filter information about these threats and create an awareness of cybersecurity threats (Sarker et al., 2020).

The situation awareness theory postulated by (Xin et al., 2019) posits that the success of a cybersecurity domain depends on its ability to obtain real-time, accurate and complete information about cybersecurity events or incidents (Menzes et al., 2016). The situation awareness model consists of situation awareness, decisions and action performance as shown in Figure 3. The

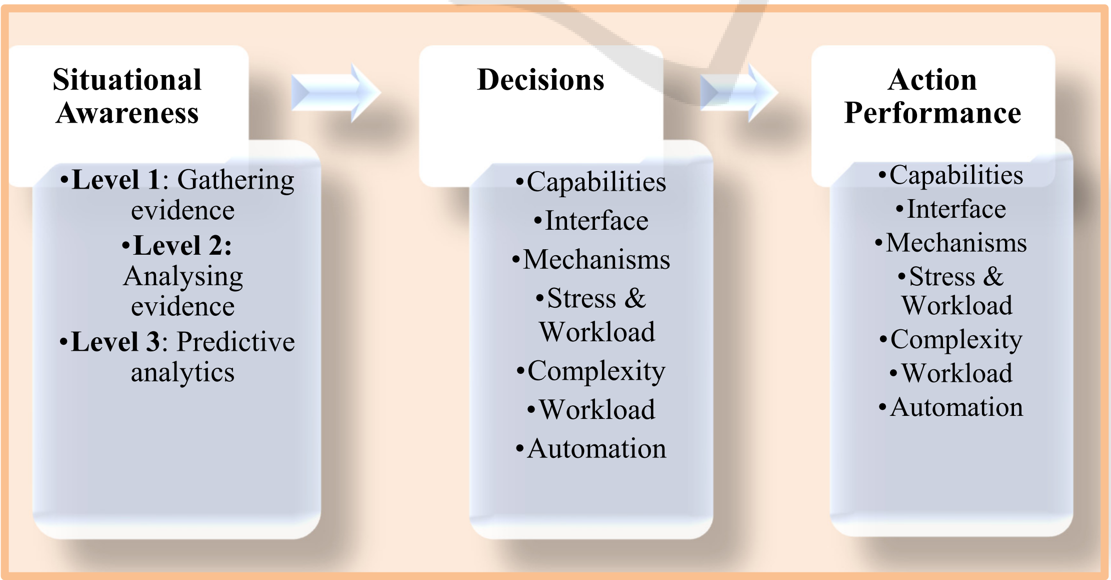


Figure 3: simplified theoretical model based on situation awareness.

transformation and expansion of the cyberspace (Cox and Wang, 2014) has rendered traditional intrusion detection and malware detection systems obsolete. Further, even the data mining models that have been used in the past are no longer sufficient for the challenges in cyber security (Cox and Wang, 2014).

A big data analytics model for cybersecurity can be evaluated on the basis of its agility and robustness (Cox and Wang, 2014). According to (Pense, 2014), Big Data is defined not only by the amount of the information that it delivers but also by its complexity and by the speed that it is analyzed and delivered.

2.5 Advances in Cloud Computing

Cloud computing is about using the internet to access someone else's software running on someone else's hardware in someone else's data center (Umamaheswari and Sujatha, 2017). Cloud Computing is essentially virtualized distributed processing, storage, and software resources and a service, where the focus is on delivering computing as a on-demand, pay-as-you-go service. The NIST Cloud computing framework states that cloud computing is made up of five essential characteristics, three service models and four deployment models (Gheyas and Abdallah, 2016), as shown on Figure 4. The five (5) essential characteristics of Cloud Computing are briefly explained follows:

1.On-demand Self-service: A consumer can unilaterally provision computing capabilities such as server time and network storage as needed automatically, without requiring human interaction with a service provider.

2.Broad Network Access: Heterogeneous client platforms available over the network come with numerous capabilities that enable provision of network access.

3.Resource Pooling: Computing resources are pooled together in a multi-tenant model depending on the consumer demand in a location independent manner.

4.Rapid Elasticity:
This is when unlimited capabilities are rapidly and elastically provisioned or purchased to quickly scale out; and rapidly released to quickly scale in.

5.Measured Service:
A transparent metering capability can be automatically controlled and optimized in cloud systems at some level of abstraction appropriate to the type of service.

Service delivery in Cloud computing comprises three (3) Cloud Service Models, namely Software-as-a-Service (SaaS), Platform-as-a-Service (PaaS) and Infrastructure-as-a-Service (IaaS). These three models are shown on Figure 5, are discussed below.

2.5.1 Software as a Service (SaaS)

The provider's applications running on a cloud infrastructure provide a capability to the consumer for use. It utilizes the Internet to deliver applications to the consumers (e.g., Google Apps, Salesforce, Dropbox, Sage X3 and office 365) (Buyya et al., 2008). This is about a wide range of applications from social to enterprise applications such as email hosting, enterprise resource planning and supply chain management. The consumer only handles minimal user specific application configuration settings. SaaS provides off-the-shelf applications offered over the internet and is the most widely used service model (Gheyas and Abdallah, 2016); (Hadi, 2015). Examples include Google Docs, Aviary, Pixlr, and the Microsoft Office Web Application.

2.5.2 Platform as a Service (PaaS)

PaaS provides to the consumer infrastructure for third-party applications. Just like in SaaS the consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage, but does have control over the deployed applications and possibly configuration settings for the application-hosting environment (Gheyas and Abdallah, 2016); (Hadi, 2015). Examples include Windows Azure, Apache Stratos, Google App Engine, CloudFoundry, Heroku, AWS (Beanstalk) and OpenShift (Buyya et al., 2008) & (Suryavanshi, 2017). PaaS supports business agility and provides an enabling environment for a consumer to run applications and services including Language, Operating System (OS), Database, Middleware and Other applications.

2.5.3 Infrastructure as a Service (IaaS)

This provisions processing, networks, storage, and other essential computing resources on which the consumer is then able to install and run arbitrary software, that can include operating systems (Virtual machines (VM), appliances, etc.) and applications (Gheyas and Abdallah, 2016); (Hadi, 2015). Common global examples include Amazon Web Services (AWS), Cisco Metapod, Microsoft Azure, Rackspace and the local ones include TelOne cloud services and Dandemutande (Buyya et al., 2008).

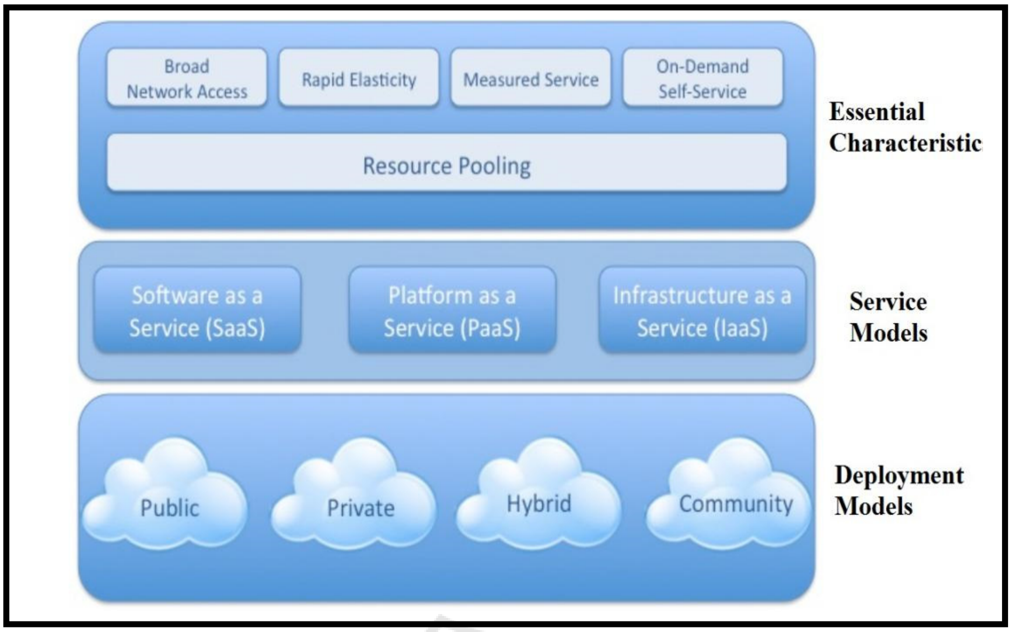


Figure 4: NIST Visual Model of Cloud Computing Definition Source: (Gheyas and Abdallah, 2016).

IaaS is a Cloud service that allows existing applications to run on its hardware. It rents out resources dynamically wherever they are needed. Services include Computer Servers, Data Storage, Firewall and Load Balancer.

3 CLOUD DEPLOYMENT MODELS

The three commonly-used cloud deployment models are private, public, and hybrid. An additional model is the community cloud which is less commonly used. In a Cloud context the term deployment basically refers to where the software is made available, in other words where it is running.

3.1 Private Cloud

The private cloud is normally either owned or exclusively used by a single organization. The services and infrastructure are permanently kept on a private network, the hardware and software are dedicated solely to the particular organisation. The major advantage of this model is the improved security as resources are not shared with others thereby allowing for higher levels of control and security (Burt et al., 2013).

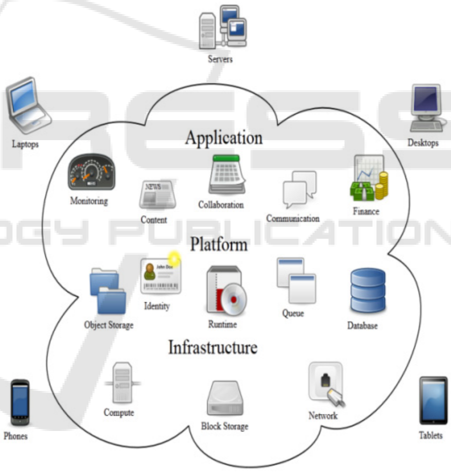


Figure 5: Cloud Computing Service Models.

3.2 Public Cloud

The cloud infrastructure is provisioned for use by the general public. The public cloud is sold to the public, as a mega-scale infrastructure, and is available to the general public. (Napanda et al., 2013) further clarifies that cloud services are provided on a subscription basis to the public. The advantages include lower costs, near-unlimited scalability and high reliability (Burt et al., 2013). Examples include Amazon (EC2), IBM's Blue Cloud, Sun Cloud, Google App Engine and Windows Azure (Marzantowicz, 2015).

3.3 Hybrid Cloud

A hybrid cloud model is a mix of two or more cloud deployment models such as private, public or hybrid (Sen and Tiwari, 2017; Fehling et al., 2014). This model requires determining the best split between the public and private cloud components. The advantages include control over sensitive data (private cloud), flexibility, ease of gradual migration (Burt et al., 2013), and data and application portability (KPMG, 2018).

3.4 Community Cloud

This model is provisioned for exclusive use by a particular community of consumers bound by shared interests (e.g., policy and compliance considerations, mission and security requirements) and third-party providers (Gheyas and Abdallah, 2016). A typical example is the U.S.-based exclusive IBM SoftLayer cloud which is dedicated for use by federal agencies only. This approach builds confidence in the platform, which cloud consumers will use to process their sensitive workloads (Marzantowicz, 2015).

3.5 Cloud Computing Benefits

Cloud computing has many benefits for the organizations and these include cost savings, scalability, anytime anywhere access, use of latest software versions, energy saving and quick rollout of business solutions. The general benefits (Kobielus, 2018) include the following (Lee, 2017):

- free capital expenditure
- accessibility from anywhere at anytime
- no maintenance headaches
- improved control over documents as files will be centrally managed
- Dynamically scalable
- Device independent
- Instant (Cost-efficient and Task-Centrism)
- Private Server Cost

The NIST Cloud Computing Definition Framework is shown below on Figure 6.

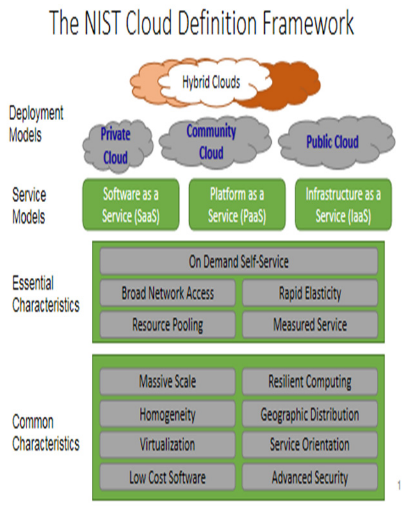


Figure 6: The NIST Cloud Computing Definition Framework.

Cloud computing leverages competitive advantage and provides improved IT capabilities. The Business benefits of Cloud Computing include the following:

- Almost zero upfront infrastructure investment
- Just-in-time Infrastructure
- More efficient resource utilization
- Usage-based costing
- Reduced time to market
- Flexibility
- Cost Reduction
- Agility
- Automatic software/hardware upgrades

The Technical Benefits of Cloud Computing are:

- Automation – “Scriptable infrastructure”
- Auto-scaling
- Proactive Scaling
- More Efficient Development lifecycle
- Improved Testability
- Disaster Recovery and Business Continuity

However, the major issues of concern and cons on Cloud Computing include the following:

- Requires a constant internet connection
- Doesn’t work well with low-speed connections
- Can be slower than using desktop software
- Features might be more limited
- Stored data might not be secure
- If the cloud loses your data, big problem
- Privacy
- Security
- Availability
- Legal Issue
- Compliance
- Performance

In conclusion the characteristics of cloud computing are leveraged through the following: Massive scale; Homogeneity; Virtualization; Resilient computing; Low cost software; Geographic distribution; Service orientation; and Advanced security technologies.

4 NETWORK FUNCTION VIRTUALIZATION

Network function virtualization (NFV) is a new paradigm to design and operate telecommunication networks. Traditionally, these networks rely on dedicated hardware-based network equipment and their functions to provide communication services. However, this reliance is becoming increasingly inflexible and inefficient, especially in dealing with traffic bursts for example during large crowd events. NFV strives to overcome current limitations by (1) implementing network functions in software and (2) deploying them in a virtualized environment. The resulting virtualized network functions (VNFs) require a virtual infrastructure that is flexible, scalable and fault tolerant.

Virtualization is basically making a virtual image or “version” of something usable on multiple machines at the same time. Virtualization technology has the drawbacks of the chance of a single point of failure of the software achieving the virtualization and the performance overhead of the entire system due to virtualization.Virtualization in general has tremendous advantages. To accommodate the needs of the industry and operating environment, to create a more efficient infrastructure – virtualization process has been modified as a powerful platform, such that the process virtualization greatly revolves around one piece of very important software.

The advantages of virtual machines are as follows:

❖ Where the physical hardware is unavailable, run the operating systems,

❖ Easier to create new machines, backup machines, etc.,

❖ Use of “clean” installs of operating systems and software for software testing

❖ Emulate more machines than are physically available,

❖ Timeshare lightly loaded systems on one host,

❖ Debug problems (suspend and resume the problem machine),

❖ Easy migration of virtual machines,

❖ Run legacy systems!

The virtualization process has been modified as a powerful platform, such that the process virtualization greatly revolves around one piece of very important software called a *hypervisor*. Thus, a VM must host an OS kernel.

Virtualization: allows the running of multiple operating systems on a single physical system and share the underlying hardware resources. Virtualization entails abstraction and encapsulation. However, Clouds rely heavily on virtualization, whereas Grids do not rely on virtualization as much as clouds. In Virtualization, a hypervisor is a piece of computer software that creates and runs virtual machines.

Instead of installing the operating system as well as all the necessary software in a virtual machine, the docker images can be easily built with a Dockerfifile since the hardware resources, such as CPU and memory, will be returned to the operating system immediately. Therefore, many new applications are programmed into containers. Cgroups allow system administrators to allocate resources such as CPU, memory, network, or any combination of them, to the running containers. This is illustrated in Figure 7 below.

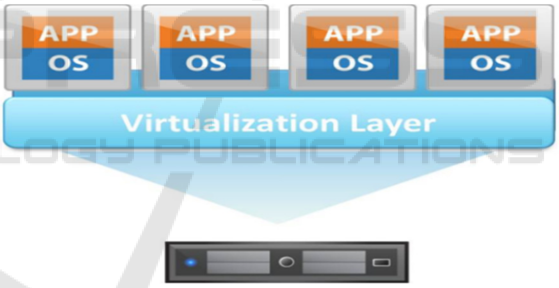


Figure 7: Architecture comparison of virtual machine Vs container.

Virtualization is the optimum way to enhance resource utilization in efficient manner. The core component of virtualization is Hypervisors. A Hypervisor is a software which provides isolation for virtual machines running on top of physical hosts. The thin layer of software that typically provides capabilities to virtual parti-tioning that runs directly on hardware. It provides a potential for virtual partitioning and responsible for running multiple kernels on top of the physical host.

Containers are different from Virtualization with respect to the following aspects:.

1. Simple:- Easy sharing of a hardware resources clean command line interface, simple REST API.
2. Fast:-Rapid provisioning, instant guest boot, and no virtualization overhead so as fast as bare metal.

3.Secure:- Secure by default, combine all available kernel security feature with AppArmor, user namespaces, SECCOMP.
4. Scalable:- The quality-of-service may be broadcast from the from a single container on a developer laptop to a container per host in a datacentre. This is also includes remote image services with Extensible storage and networking.
5. Control groups (cgroups) :- This is a kernel-provided mechanism for administration, grouping and tracking through a virtual filesystem.
Docker containers share the operating system and important resources, such as depending libraries, drivers or binaries, with its host and therefore they occupy less physical resources.

5 RESEARCH METHODOLOGY

5.1 Presentation of the Methodology

The Pragmatism paradigm was used in this research and this is intricately related to the Mixed Methods Research (MMR) .

5.1.1 Research Approach and Philosophy

The researcher adopts a qualitative approach in form of focus group discussion to research. Since the analysis is done to establish differences in data analytics models for cybersecurity without the necessity of quantifying the analysis (Iafrate, 2015) .

The researcher adopts a postmodern philosophy to guide the research. Firstly the researcher notes that the definition, scope and measurement of cybersecurity differs between countries and across nations (Moorthy et al., 2014). Further, the post-modern view is consistent with descriptive research designs which seek to interpret situations or models in their particular contexts (Vadapalli, 2020).

5.1.2 Research Design and Methods

The researcher adopts a descriptive research design since the intention is to systematically describe the facts and characteristics of big data analytics models for cybersecurity. The purpose of the study is essentially an in-depth description of the models (Iafrate, 2015).

A case study research method was adopted in this study. In this respect each data analytics model for cybersecurity is taken as a separate case to be investigated in its own separate context (Vadapalli, 2020). Prior research has tended to use case studies in relation to the study of cybersecurity (Moorthy et al., 2014). However, the researcher develops a control case that accounts for an ideal data analytics model for cybersecurity for comparative purposes.

5.2 Population and Sampling

5.2.1 Population

The research population for the purpose of this study consists of all data analytics models for cybersecurity that have been proposed and developed in literature, journals, conference proceedings and working papers. This is consistent with previous research which involves a systematic review of literature (Gheyas and Abdallah, 2016).

5.2.2 Sample

The researcher identified two data analytics models or frameworks from a review of literature and the sample size of 8. Eight participants in total were interviewed. However, while this may be limited data, it will be sufficient for the present needs of this study. Research in future may review more journals to identify more data analytics models which can be applied to cybersecurity.

5.3 Sources and Types of Data

The researcher uses secondary data in order to investigate the application of data analytics models in cybersecurity.

5.4 Model for Analysis

In analyzing the different data analytics models for cybersecurity the researcher makes reference to the characteristics of an ideal data analytics model for cybersecurity. In constructing an ideal model, the researcher integrates various literature sources. The basic framework for big data analytics model for cybersecurity consists of three major components which are big data, analytics, and insights (Cox and Wang, 2014). However, a fourth component may be identified as prediction (or predictive analytics) (Gheyas and Abdallah, 2016). This is depicted in Figure 8 below:

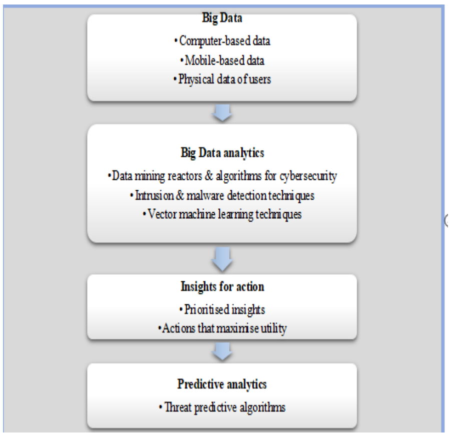

Big Data
•Computer-based data
•Mobile-based data
•Physical data of users
Big Data analytics
•Data mining reactors & algorithms for cybersecurity
•Intrusion & malware detection techniques
•Vector machine learning techniques
Insights for action
•Prioritised insights
•Actions that maximise utility
Predictive analytics
•Threat predictive algorithms

Figure 8: Big data analytics model for cybersecurity.

5.5 Big Data Analytics

The address the concerns of big data about cybersecurity, more robust big data analytics models for cybersecurity have been developed in data mining techniques and machine learning (Cox and Wang, 2014). Big data analytics employ data mining reactors and algorithms, intrusion and malware detection techniques and vector machine learning techniques for cybersecurity (Cox and Wang, 2014). However, it has been observed that adversarial programs have tended to modify their behavior by adapting to the reactors and algorithms designed to detect them (Cox and Wang, 2014). Further, intrusion detection systems are faced with challenges such as unbounded patterns, data nonstationarity, uneven time lags, individuality, high false alarm rates, and collusion attacks (Gheyas and Abdallah, 2016). This necessitates a multi-layered and multi-dimensional approach to big data analytics for cybersecurity (Hammond, 2015), (Sarker et al., 2020). In other words an effective big data analytics model for cybersecurity must be able to detect intrusions and malware at every layer in the cybersecurity framework.

5.6 Predictive Analytics

Predictive analytics refer to the application of a big data analytics model for cybersecurity to derive, from current cybersecurity data, the likelihood of a cybersecurity event occurring in future (Gheyas and Abdallah, 2016). In essence, a data analytics model for cybersecurity should be able to integrate these components if it is to be effective in its major functions of gathering big data about cybersecurity, analyzing big data about cybersecurity threats, providing actionable insights and predicting likely future cybersecurity incidents.

5.7 Validity and Reliability

The researcher solicited comments from peers on the emerging findings and also feedback to clarify the biases and assumptions of the researcher to ensure internal validity of the study (Vadapalli, 2020).

5.8 Possible Outcomes

The expected accuracy rate for the research should be according to Table 3 below, which shows the international benchmark.

Table 3: Comparative Detection accuracy rate (%).

Classifier	Detection Accuracy (%)	Time taken to build the Model in seconds	False Alarm rate (%)
Decision Trees (J48)	81.05	**	**
Naive Bayes	76.56	**	**
Random Forest	80.67	**	**
SVM	69.52	**	**
AdaBoost	90.31	**	3.38
Mutlinomal Naive Bayes + N2B	38.89	0.72	27.8
Multinomal Naive Bayes updateable + N2B	38.94	1.2	27.9
Discriminative Multinomal Bayes + PCA	94.84	118.36	4.4
Discriminative Multinomal Bayes + RP	81.47	2.27	12.85
Discriminative Multinomal Bayes + N2B	96.5	1.11	3.0

6 ANALYSIS AND RESEARCH OUTCOMES

6.1 Overview

The analysis of possible attacks on an intrusion network is shown on Figure 9 below.

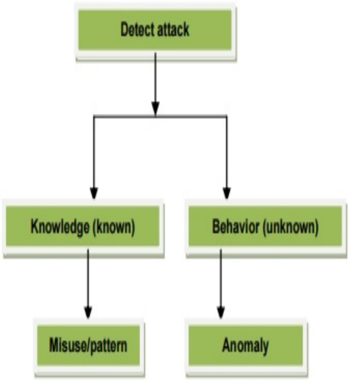


Figure 9: Analysis of Attack (Source: (Hashem et al., 2015)).

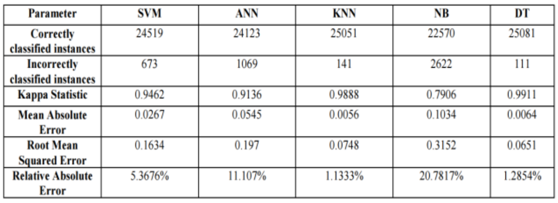
(Moorthy et al., 2014) highlighted the use of Machine Learning (ML), Neural Network and Fuzzy Logic to detect attacks on private networks on the different Artificial Intelligence (AI) techniques. It is not technically feasible to develop a perfect sophisticated Intrusion Detection System, since the majority of IDS are signature based.

The IDS is divided into either as a Host IDS (HIDS) or as a Network IDS (NIDS). Analysis of the network traffic can be handled by a NIDS which distinguishes the unlicensed, illegitimate and anomalous behavior on the network. Packets traversing through the network should generally be captured by the IDS using network taps or span port in order to detect and flag any suspicious activity (Moorthy et al., 2014)). Anomalous behavior on the specific device or malicious activity can be effectively detected by a device specific IDS. The vulnerability of networks and susceptibility to cyber attacks is exacerbated by the use of wireless technology (Hammond, 2015).

The gross inadequacies of classical security measures have been overtly exposed. Therefore, effective solutions for a dynamic and adaptive network defence mechanism should be determined. Neural networks can provide better solutions for the representative sets of training data (Hammond, 2015). (Hammond, 2015) argues for the use of ML classification problems solvable with supervised or semi-supervised learning models for the majority of the IDS. However, the one major limitation of the work done by (Hammond, 2015) is on the informational structure in cybersecurity for the analysis of the strategies and the solutions of the players.

Intrusion attack classification requires optimization and enhancement of the efficiency of data mining techniques. The pros and cons of each algorithm using the NSL-KDD dataset are shown on Table 4 below.

Table 4: Performance of Support Vector Machines, Artificial Neural Network, K-Nearest Neighbour, Naive-Bayes and Decision Tree Algorithms.



Parameter	SVM	ANN	KNN	NB	DT
Correctly classified instances	24519	24123	25051	22570	25081
Incorrectly classified instances	673	1069	141	2622	111
Kappa Statistic	0.9462	0.9136	0.9888	0.7906	0.9911
Mean Absolute Error	0.0267	0.0545	0.0056	0.1034	0.0064
Root Mean Squared Error	0.1634	0.197	0.0748	0.3152	0.0651
Relative Absolute Error	5.3676%	11.107%	1.1333%	20.7817%	1.2854%

An intrusion detection system determines if an intrusion has occurred, and so monitors computer systems and networks, and the IDS raises an alert when necessary (Truong et al., 2020). However, (Truong et al., 2020) addressed the problems of Anomaly Based Signature (ABS) which reduces false positives by allowing a user to interact with the detection engine and raising classified alerts. The advantages and disadvantages of ABSs and SBSs are summarised on table, Table 5, below.

Table 5: Advantages and disadvantages of ABSs and SBSs models (Source: (Truong et al., 2020)).

Detection Model	Advantages	Disadvantages
Signature-based	Low false positive rate Does not require training Classified alerts	Cannot detect new attacks Requires continuous updates Tuning could be a thorny task
Anomaly-based	Can detect new attacks Self-learning	Prone to raise false positives Black-box approach Unclassified alerts Requires initial training

An IDS must keep up track of all the data, networking components and devices involved. Additional requirements must be met when developing a cloud-based intrusion detection system due to its complexity and integrated services.

6.2 Support Vector Machine

Support Vector Machine is a classification artificial intelligence and machine learning algorithm with a set containing of points of two types in X dimensional place. Support vector machine generates a (X—1) dimensional hyperplane for separating these points into two or more groups using either linear kernel or non-linear kernel functions (Menzes et al., 2016). Kernel functions provides a method for polynomial, radial and multi-layer perception classifiers such as classification of bank performance into four clusters of strong, satisfactory, moderate and poor performance. The class of bank performance is defined by the function

$$Performance\ class = f\left(\vec{x}.\vec{w}\right) = f\left(\sum_j x_j w_j\right)$$

Where $\vec{x}$ is the input vector to the support vector classifier and $\vec{w}$ is the real vector of weights and f is the function that translates the dot product of the input and real vector of weights into desired classes of bank performance. $\vec{w}$ Is learned from the labeled training data set.

6.3 KNN Algorithm

The K-NN algorithm is a non-parametric supervised machine learning technique that endeavors to classify a data point from given categories with the support of the training dataset (Menzes et al., 2016). Predictions are performed for a new object (y) by searching through the whole training dataset for the K most similar instances or neighbors. The algorithm does this by calculating the Euclidean distance as follows:

$$d(x,y) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2}$$

Where $d(x,y)$ is the distance measure for finding the similarity between new observations and training cases and then finding the k-closest instance to the new instance. Variables are standardized before calculating the distance since they are measured in different units. Standardization is performed by the following function:

$$X_s = \frac{X - mean}{s.d}$$

Where X_s is the standardized value, X is the instance measure, mean and s.d are the mean and standard deviation of instances. Lower values of K are sensitive to outliers and higher values are more resilient to outliers and more voters are considered to decide the prediction.

6.4 Multi Linear Discriminant Analysis (LDA)

The Linear Discriminant Analysis is a dimensionality reduction technique. Dimensionality reduction is the technique of reducing the amount of random variables under consideration through finding a set of principal variables (Menzes et al., 2016) which is also known as course of dimensionality. The LDA calculates the separability between n classes also known as between-class variance. Let D_b be the distance between n classes.

$$D_b = \sum_{i=1}^{g} N_i\left(\overline{x_i} - \overline{x}\right)\left(\overline{x_i} - \overline{x}\right)^T$$

Where $\overline{x}$ the overall is mean, $\overline{x_i}$ and N_i are the sample mean and sizes of the respective classes. The within-class variance is then calculated, which is the distance between mean and the sample of every class. Let S_y be the within class variance.

$$S_y = \sum_{t=1}^{g}(N_i - 1)S_i = \sum_{t=1}^{g}\sum_{j}^{N_i}\left(X_{i,j} - \overline{X_i}\right)\left(X_{i,j} - \overline{X_i}\right)^2$$

The final procedure is to then construct the lower dimensional space for maximization of the seperability between classes and the minimization of within class variance. Let P be the lower dimensional space.

$$\mathrm{P} = arg_p \max \frac{|P^T D_b \mathrm{P}|}{|P^T S_y P|}$$

The LDA estimates the probability that a new instance belongs to every class. Bayes Theorem is used to estimate the probabilities. For instance, if the output of the class is (a) and the input is (b) then

$$P(Y = x \mid B = b) = (P \mid a * fa(b)) / \sum(P \mid a * f \mid (b))$$

P|a is the prior probability of each class as observed in the training dataset and f(b) is the estimated probability of b belonging to the class, f(b) uses the Gaussian distribution function to determine whether b belongs to that particular class.

6.5 Random Forest Classifier

The Random Forest classifier is an ensemble algorithm used for both classification and regression problems. It creates a set of decision trees from a randomly selected subset of the training set (Menzes et al., 2016). The tree with higher error rates are given low weight in comparison to other trees increasing the impact of trees with low error rate.

6.6 Variable Importance

Variable importance was implemented using the

Boruta algorithm to improve model efficiency. The Boruta algorithm endeavors to internment all the key, interesting features existing in the dataset with respect to an outcome variable. The diagram below shows that net profit is the most significant feature, followed by ROA, total assets, ROE and other variables depicted below in Figure 10.

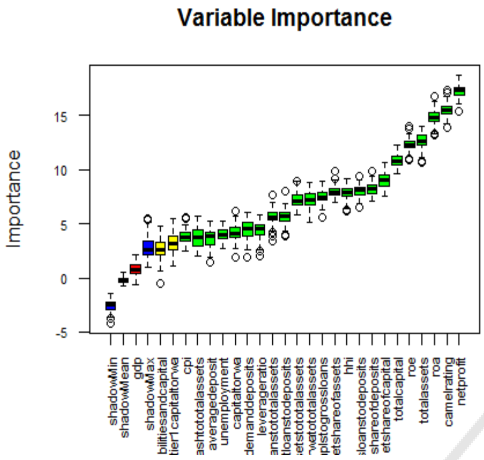


Figure 10: Boruta Algorithm important features.

The next procedure was fitting these variable into our algorithms and hence evaluating their performance using the metrics discussed in the models section. The Boruta algorithm also clusters banks on important variable as shown below in Figure 11 for effective risk management and analysis.

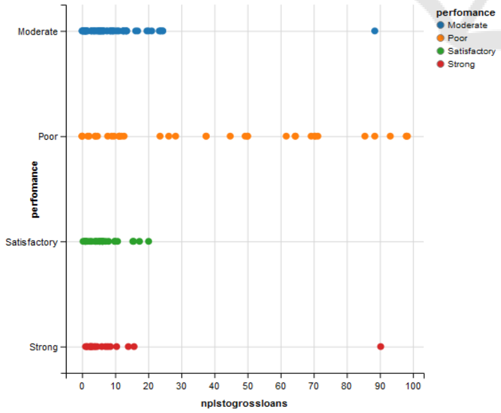


Figure 11: Boruta algorithm clustering banks based on non-performing loans.

6.7 Model Results

Before we discuss the results of our models. It is imperative to discuss the distribution of our dataset. We classify bank performance into four classes which are strong, satisfactory, moderate and poor performing banks. A strongly performing bank is the one with incredible CAMELS indicators. Its profitability indicators are high, the management quality is top of the class, less sensitive to market movements with a high quality asset base. A satisfactory bank is the one with acceptable but not outstanding performance.

Our dataset comprises thousands of records from banking institutions returns. The distribution of performance classes is shown on the diagram below. We can see that strong banks comprise of 12.9%, satisfactory banks 15.1%, moderate banks 47.5% and poor banks 24.5%. Figure 12 visualizes the effectiveness of Boruta algorithm in determining the most important variables that determines the condition of a bank.

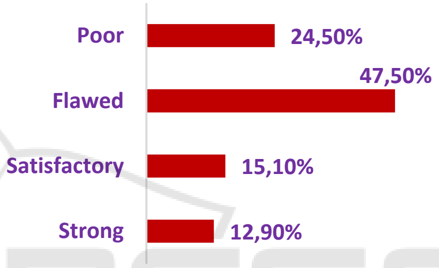


Figure 12: Distribution of the big dataset.

6.8 Classification and Regression Trees (CART)

Table 6 below shows the performance results of our CART algorithm in predicting bank failure on the training set. The algorithm's level of accuracy on the training dataset was 82.8%. The best tune or complexity parameter of our optimal model was 0.068. The Kappa statistic was 75% envisaging that our classifier was effective as also shown with the Kappa SD of 0.07 in the classification of bank categories. On the test dataset, the algorithm achieved an accuracy level of 92.5% and a kappa of 88.72%. The algorithm only misclassified 2 instance as moderate and 1 as satisfactory.

Table 6: CART model performance.

Complexity Parameter	Accuracy	Kappa	AccuracySD	KappaSD
0.06849315	0.8275092	0.7519499	0.04976459	0.07072572
0.15753425	0.7783150	0.6683229	0.07720896	0.14039942
0.42465753	0.5222344	0.1148591	0.08183351	0.18732422

The accuracy of the CART model based on the complexity parameters of different test runs is shown on Figure 13 below. The complexity parameter or the best tune parameter of 0.068 optimized the model performance.

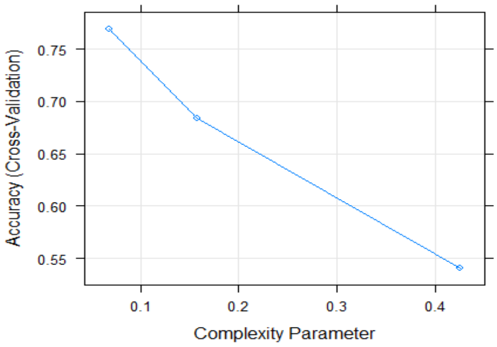


Figure 13: CART accuracy curve.

6.9 Support Vector Machine

The accuracy level of the SVM model on the training dataset was 79.1% in predicting bank solvency as shown in table 7. The best tune sigma and cost values of our highly performing model where 0.05 and 1 as shown on Figure 14 below. The Kappa statistic and the Kappa SD where 67.9% and 0.13 respectively. On the test dataset, the algorithm achieved an accuracy level of 92.5% and a kappa of 88.54%. The algorithm only misclassified 3 instance as moderate in comparison to the CART algorithm.

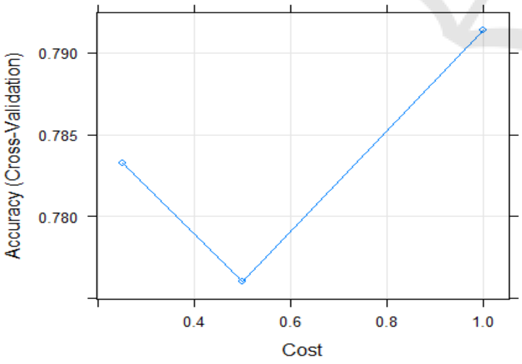


Figure 14: SVM accuracy curve.

Table 7: Support Vector Machine performance.

sigma	c	Accuracy	Kappa	AccuracySD	KappaSD
0.050398	0.25	0.783223	0.678536	0.095598	0.140312
0.050398	0.50	0.776007	0.661354	0.087866	0.132552
0.050398	1.00	0.791391	0.678694	0.080339	0.126466

6.10 Linear Discriminant Algorithm

Table 8: Linear Discriminant algorithm performance.

Accuracy	Kappa	AccuracySD	KappaSD
0.8042399	0.7038131	0.1016816	0.159307

On the training dataset, the LDA achieved an accuracy level of 80% as in table 8. The Kappa statistic and the Kappa SD where 70% and 0.16 respectively. On the test dataset, the algorithm achieved an accuracy level of 90% and a kappa of 84.64%. The algorithm only misclassified 4 instance as moderate whose performance is poor in comparison to the CART algorithm.

6.11 K-Nearest Neighbor

Table 9: K-NN algorithm performance.

K	Accuracy	Kappa	AccuracySD	KappaSD
5	0.5988645	0.3698931	0.1280376	0.2158109
7	0.6268864	0.4072928	0.1564920	0.2703504
9	0.6621978	0.4715556	0.1747903	0.2881390

The level of accuracy on the training dataset was 66.2%. The best tune parameter for our model was k=9 or 9 neighbors as shown on the accuracy curve in Figure 15 below. The Kappa statistic and the Kappa SD where 47.2% and 0.17 respectively. On the test dataset, the algorithm achieved an accuracy level of 67.5% and a kappa of 49%. The algorithm was not highly effective in classifying bank performance in comparison to other algorithms.

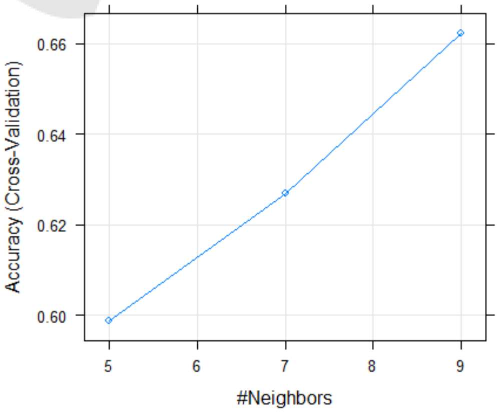


Figure 15: K-NN confusion accuracy graph.

6.12 Random Forest

Table 10: Random Forest performance.

mtry	Accuracy	Kappa	AccuracySD	KappaSD
2	0.8272527	0.7421420	0.10396454	0.15420079
14	0.8554212	0.7829891	0.06069716	0.09303130
16	0.8482784	0.7718935	0.06455248	0.09881991

On the training set, the accuracy of our random forest was 85.5% as designated in table 10. The best tune parameter for our model was the mtry of 14 which is the number of randomly selected predictors in constructing trees as shown on Figure 16. The Kappa statistic and the Kappa SD where 78.3% and 0.09 respectively. On the test dataset, the algorithm achieved an accuracy level of 96% and a kappa of 96%. The algorithm was highly effective in classifying bank performance in comparison to all algorithms.

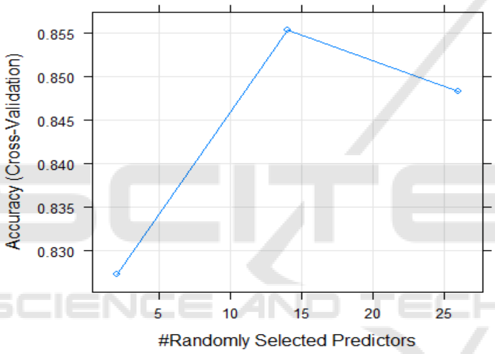


Figure 16: Random forest accuracy graph.

6.13 Challenges and Future Direction

As number of banking activities increase, also implies that the data submission to the Reserve Bank continues to grow exponentially. This challenging situation in combination with advances in machine learning (ML) and artificial intelligence (AI) presents unlimited opportunities to apply neural network-based deep learning (DL) approaches to predict Zimbabwean Bank's solvency. Future work will focus on identifying more features that could possibly lead to poor bank performance and incorporate these in our models to develop a robust early warning supervisory tool based on big data analytics, machine learning and artificial intelligence.

The researcher analyses the two models that have been proposed in literature with reference to an ideal data analytics model for cybersecurity presented in section 3.

6.13.1 Model 1: Experimental/ Prototype Model

In the first case the researcher makes reference to the model presented in (Petrenko and Makovechuk, 2020) which although developed in the context of the public sector can be applied to the private sector organizations. Table 11 below summarizes the main characteristics of the experimental model.

Software and Hardware Complex (SHC): Warning-2016

Table 11: Experimental big data analytics model for cybersecurity.

MODEL ATTRIBUTES	DESCRIPTION
HBase working on HDFS (Hadoop Distributed File System)	• HBase, a non-relational database, facilitates analytical and predictive operations • Enables users to assess cyber-threats and the dependability of critical infrastructure
Analytical data processing **module**	• Processes large amounts of data, interacts with standard configurations servers and is implemented at C language • Special interactive tools (based on JavaScript/ CSS/ DHTML) and libraries (for example jQuery) developed to work with content of the proper provision of cybersecurity
Special interactive tools and libraries	• Interactive tools based on JavaScript/ CSS/ DHTML • Libraries for example jQuery developed to work with content for • Designed to ensure the proper provision of cybersecurity
Data store for example (MySQL)	• Percona Server with the ExtraDB engine • DB servers are integrated into a multi-master cluster using the Galera Cluster.
Task queues and data caching	• Redis
Database servers balancer	• Haproxy
Web server	• nginx, involved PHP-FPM with APC enabled
HTTP requests balancer	• DNS (Multiple A-records)
Development of special client applications running Apple iOS	• Programming languages are used: Objective C, C++, Apple iOS SDK based on Cocoa Touch, CoreData, and UIKit.
Development of applications running Android OS	• Google SDK
Software development for the web platform	• PHP and JavaScript.
Speed of the service and protection from DoS attacks	• CloudFare (through the use of CDN)

(Source: (Petrenko and Makovechuk, 2020)).

The proposed model, it is to be noted was demonstrated to be effective in integrating big data analytics with cybersecurity in a cost effective way (Petrenko and Makovechuk, 2020).

6.13.2 Model 2: Cloud Computing/Outsourcing

The second model involves an organization outsourcing its data to a cloud computing service provider. Cloud computing service providers usually have advanced big data analytics models, with advanced detection and prediction algorithms and better state of the art cybersecurity technologies and better protocols because they specialize in data and networks. However, it is to be noted that cloud computing service providers are neither exempt nor immune from cyber-threats and attacks(Mazumdar and Wang, 2018).

6.13.3 Application of Big Data Analytics Models in Cybersecurity

The researcher demonstrated by identifying the characteristics of an effective data analytics model, the ideal model, that it is possible to evaluate different models. While the review of literature showed that institutions and countries adopt different big data analytics models for cybersecurity, the researcher also demonstrated that beside the unique requirements these models share major common characteristics for example reactors and detection algorithms are usually present in every model but differ in terms of complexity.

The first experimental or prototype model involves the design, and implementation of a prototype by an institution and the second model involves the use serviced provided by cloud computing companies. By applying such analytics to big data, valuable information can be extracted and exploited to enhance decision making and support informed decisions.

7 CONCLUSION

The main characteristic of Machine Learning is the automatic data analysis of large data sets and production of models for the general relationships found among data. Big data analytics is not only about the size of data but also clinches on volume, variety and velocity of data. The information that was evaluated in Big Data Analytics includes a mixer of unstructured and semi-structured data, for instance, social media content, mobile phone records, web server logs, and internet click stream data.

A Cloud computing setting was added which has advanced big data analytics models and advanced detection and prediction algorithms to strengthen the cybersecurity system. IoT requires both cloud computing environment to handle its data exchange and processing; and the use of artificial intelligence for data mining and data analytics.

Big data analytics makes use of analytic techniques such as data mining, machine learning, artificial learning, statistics, and natural language processing. In an age of transformation and expansion in the Internet of Things , cloud computing services and big data, cyber-attacks have become enhanced and complicated , and therefore cybersecurity events become difficult or impossible to detect using traditional detection systems.

As a result, there is an imperative for security network administrators to be more flexible, adaptable, and provide robust cyber defense systems in real-time detection of cyber threats.The key problem is to evaluate Machine Learning and Big Data Analytics paradigms for use in Cybersecurity.

The traditional examples of machine learning algorithms include Linear regression, Logistic regression, Linear discriminant analysis, classification and regression trees, Naïve bayes, K-Nearest Neighbour , Kmeans clustering Learning Vector Quantization , Support Vector Machines , Random Forest, Monte Carlo, Neural networks and Q-learning.

REFERENCES

Berman, D.S., Buczak, A.L., Chavis, J.S., and Corbett, C.L. (2019). "Survey of Deep Learning Methods for Cyber Security", *Information* 2019, *10*, 122; doi:10.3390/info10040122.

Sarker, I. H., Kayes, A. S. M., Badsha, S., Alqahtani, H., Watters, P., & Ng, A. (2020). Cybersecurity data science: an overview from machine learning perspective. *Journal of Big Data*. https://doi.org/10.1186/s40537-020-00318-5

Bringas, P.B., and Santos, I., (2010). Bayesian Networks for Network Intrusion Detection, Bayesian Network, Ahmed Rebai (Ed.), ISBN: 978-953-307-124-4, InTech, Available from: http://www.intechopen.com/books/bayesian-network/bayesian-networks-for-network-intrusion-detection

Truong, T.C; Diep, Q.B.; & Zelinka, I. (2020). Artificial Intelligence in the Cyber Domain: Offense and Defense. Symmetry 2020, 12, 410.

Stefanova, Z.S., (2018). "Machine Learning Methods for Network Intrusion Detection and Intrusion Prevention

Systems", Graduate Theses and Dissertations, 2018, https://scholarcommons.usf.edu/etd/7367

Proko, E., Hyso, A., and Gjylapi, D. (2018). Machine Learning Algorithms in Cybersecurity, http://www.CEURS-WS.org/Vol-2280/paper-32.pdf

Mazumdar, S & Wang J (2018). Big Data and Cyber security: A visual Analytics perspective in S. Parkinson et al (Eds), Guide to Vulnerability Analysis for Computer Networks and Systems.

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Ullah Khan, S. (2015). The rise of "big data" on cloud computing: Review and open research issues. In *Information Systems*. https://doi.org/10.1016/j.is.2014.07.006

Siti Nurul Mahfuzah, M., Sazilah, S., & Norasiken, B. (2017). An Analysis of Gamification Elements in Online Learning To Enhance Learning Engagement. *6th International Conference on Computing & Informatics*.

Moorthy, M., Baby, R. & Senthamaraiselvi, S., 2014. An Analysis for Big Data and its Technologies. *International Journal of Computer Science Engineering and Technology(IJCSET),* 4(12), pp. 413-415.

Cox, R. & Wang, G., 2014. Predicting the US bank failure: A discriminant analysis. *Economic Analysis and Policy,* Issue 44.2, pp. 201-211.

Hammond, K., 2015. *Practical Artificial Intelligence For Dummies®, Narrative Science Edition.* Hoboken, New Jersey: John Wiley & Sons, Inc.

Yang, C., Yu, M., Hu, F., Jiang, Y., & Li, Y. (2017). Utilizing Cloud Computing to address big geospatial data challenges. *Computers, Environment and Urban Systems*. https://doi.org/10.1016/j.compenvurbsys.2016.10.010

Jiang, W., Wang, L., & Lin, H. (2016). The role of cognitive processes and individual differences in the relationship between abusive supervision and employee career satisfaction. *Personality and Individual Differences*. https://doi.org/10.1016/j.paid.2016.04.088

Fernando, J. I., & Dawson, L. L. (2009). The health information system security threat lifecycle: An informatics theory. *International Journal of Medical Informatics*. https://doi.org/10.1016/j.ijmedinf.2009.08.006

Menzes, F.S.D., Liska, G.R., Cirillo, M.A. and Vivanco, M.J.F. (2016) Data Classification with Binary Response through the Boosting Algorithm and Logistic Regression. Expert Systems with Applications, 69, 62-73. https://doi.org/10.1016/j.eswa.2016.08.014

Petrenko, S A & Makovechuk K A (2020). Big Data Technologies for Cybersecurity.

Pense (2014), Pesquisa Nacional de Saude do Escolar, Rio de Janeiro, RJ - Brazil.

Xin, Y., Kong, L., Liu, Z., Chen, Y., Li, Y., Zhu, H., Gao, M., Hou, H., & Wang, C. (2018). Machine Learning and Deep Learning Methods for Cybersecurity. *IEEE Access*, *6*, 35365–35381. https://doi.org/10.1109/ACCESS.2018.2836950

Umamaheswari, K., and Sujatha, S., (2017). Impregnable Defence Architecture using Dynamic Correlation-based Graded Intrusion Detection System for Cloud, Defence Science Journal, Vol. 67, No. 6, November 2017, pp. 645-653, DOI : 10.14429/dsj.67.11118.

Gheyas, I. A. & Abdallah, A. E. (2016). Detection and prediction of insider threats to cyber security: A systematic Literature Review and Meta-Analysis., Big Data Analytics (2016) 1:6.

Buyya, R., Yeo, C. S., & Venugopal, S. (2008). Market-oriented cloud computing: Vision, hype, and reality for delivering IT services as computing utilities. *Proceedings - 10th IEEE International Conference on High Performance Computing and Communications, HPCC 2008*. https://doi.org/10.1109/HPCC.2008.172

Hadi, J., (2015) 'Big Data and Five V'S Characteristics', *International Journal of Advances in Electronics and Computer Science*, (2), pp. 2393–2835.

Suryavanshi, A., (2017), "Magnesium oxide nanoparticle-loaded polycaprolactone composite electrospun fiber scaffolds for bone–soft tissue engineering applications: in-vitro and in-vivo evaluation", 2017 Biomed. Mater. 12 055011, https://iopscience.iop.org/article/10.1088/1748-605X/aa792b/pdf

Burt, D., Nicholas, P., Sullivan, K., & Scoles, T. (2013). Cybersecurity Risk Paradox. *Microsoft SIR*.

Napanda, K., Shah, H., and Kurup, L., (2015). Artificial Intelligence Techniques for Network Intrusion Detection, International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, IJERTV4IS110283 www.ijert.org, Vol. 4 Issue 11, November-2015.

Marzantowicz, (2015), Corporate Social Responsibility of TSL sector: attitude analysis in the light of research, „Logistyka" 2014, No. 5, pp. 1773—1785.

Sen and Tiwari, (2017). Port sustainability and stakeholder management in supply chains: A framework on resource dependence theory, The Asian Journal of Shipping and Logistics, No. 28 (3): 301-319.

Fehling, C., Leymann, F., Retter, R., Schupeck, W., & Arbitter, P. (2014). Cloud Computing Patterns. In *Cloud Computing Patterns*. https://doi.org/10.1007/978-3-7091-1568-8

KPMG (2018) , Clarity on Cybersecurity. Driving growth with confidence.

Kobielus, J., (2018). Deploying Big Data Analytics Applica- tions to the Cloud: Roadmap for Success. Cloud Standards Customer Council

Lee, J. (2017). Hacking into China's cybersecurity law, In: IEEE International Conference on Distributed Computing Systems (2017).

Iafrate, F., (2015), From Big Data to Smart Data, ISBN: 978-1-848-21755-3 March, 2015, Wiley-ISTE, 190 Pages.

Pavan Vadapalli, (2020). "AI vs Human Intelligence: Difference Between AI & Human Intelligence", 15th September, 2020, https://www.upgrad.com/blog/ai-vs-human-intelligence/

Almutairi, A., (2016). Improving intrusion detection systems using data mining techniques, Ph.D Thesis, Loughborough University, 2016.